Using self-organising maps to explore ozone profile validation results – SCIAMACHY limb compared to ground-based lidar observations

J. A. E. van Gijsel, R. Zurita-Milla, P. Stammes, Sophie Godin-Beekmann, Thierry Leblanc, Marion Marchand, I. S. Mcdermid, K. Stebel, W. Steinbrecht, D. P. J. Swart

Atmos. Meas. Tech., 8, 1951–1963, 2015
www.atmos-meas-tech.net/8/1951/2015/
doi:10.5194/amt-8-1951-2015

Atmospheric Measurement Techniques
Open Access

Using self-organising maps to explore ozone profile validation results – SCIAMACHY limb compared to ground-based lidar observations

J. A. E. van Gijsel[1], R. Zurita-Milla[2], P. Stammes[1], S. Godin-Beekmann[3], T. Leblanc[4], M. Marchand[3], I. S. McDermid[4], K. Stebel[5], W. Steinbrecht[6], and D. P. J. Swart[7]

[1]Royal Netherlands Meteorological Institute (KNMI), De Bilt, the Netherlands
[2]University of Twente, Enschede, the Netherlands
[3]LATMOS IPSL CNRS/UPMC/UVSQ, Paris, France
[4]NASA/JPL/California Institute of Technology, Wrightwood, CA, USA
[5]Norwegian Institute for Air Research (NILU), Oslo, Norway
[6]German Weather Service (DWD), Hohenpeißenberg, Germany
[7]National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands

Correspondence to: J. A. E. van Gijsel (anne.van.gijsel@knmi.nl)

Received: 25 March 2014 – Published in Atmos. Meas. Tech. Discuss.: 30 April 2014
Revised: 31 March 2015 – Accepted: 3 April 2015 – Published: 6 May 2015

Abstract. Traditional validation of atmospheric profiles is based on the intercomparison of two or more data sets in predefined ranges or classes of a given observational characteristic such as latitude or solar zenith angle. In this study we trained a self-organising map (SOM) with a full time series of relative difference profiles of SCIAMACHY limb v5.02 and lidar ozone profiles from seven observation sites. Each individual observation characteristic was then mapped to the obtained SOM to investigate to which degree variation in this characteristic is explanatory for the variation seen in the SOM map. For the studied data sets, altitude-dependent relations for the global data set were found between the difference profiles and studied variables. From the lowest altitude studied (18 km) ascending, the most influencing factors were found to be longitude, followed by solar zenith angle and latitude, sensor age and again solar zenith angle together with the day of the year at the highest altitudes studied here (up to 45 km). After accounting for both latitude and longitude, residual partial correlations with a reduced magnitude are seen for various factors. However, (partial) correlations cannot point out which (combination) of the factors drives the observed differences between the ground-based and satellite ozone profiles as most of the factors are inter-related. Clustering into three classes showed that there are also some local dependencies, with for instance one cluster having a much stronger correlation with the sensor age (days since launch) between 36 and 42 km. The proposed SOM-based approach provides a powerful tool for the exploration of differences between data sets without being limited to a priori defined data subsets.

1 Introduction

Accurate knowledge on the quality and stability of long-term measurements is required for time series trend analysis as well as for merging multiple data sets (Nair, 2012). Remote sensing products must therefore be compared and/or validated with independent measurements of known quality (as determined by other data sources). In the case of satellite-based atmospheric columns and profiles, this validation data source is usually formed by acquisitions from other satellite sensors (e.g. Nazaryan et al., 2007; Boersma et al., 2008), ground-based and/or in situ observers (e.g. Herman et al., 2009; van Gijsel et al., 2010; Takele Kenea et al., 2013), the combination of both (e.g. Adams et al., 2012; Stiller et al., 2012; Wetzel et al., 2013) or with the additional inclusion of model data (e.g. Lamsal et al., 2010; Zhang et al., 2010).

Traditionally, data validation and intercomparisons are made for predefined classes or ranges of possibly correlated variables which are then studied for inter-class differences to determine limitations in the retrieval scheme. In atmospheric validation studies, this usually comes down to dividing the global data set into various latitude ranges, splitting observation characteristics such as solar or stellar zenith angle and viewing angle into a few groups, studying secondary retrieval output (uncertainty estimates, processing and cloud flags, goodness-of-fit measures) and occasionally adding other data (e.g. input used in the retrieval like temperature, difference in equivalent latitude). Such a procedure has various limitations. To start with, it requires a priori knowledge, or a substantial amount of testing, on how to divide each variable (information source) into classes. Moreover, the need to have a group of classes is a limitation by itself as there might be a gradual transition from one extreme to the other and dependencies on multiple (correlated) variables further complicate the analysis procedure.

Here we will present an alternative approach to data intercomparison that traces down possible explanatory variables and patterns associated with the differences found in the data sets that are being compared and that does not require a priori grouping of variables. The approach is based on the usage of self-organising maps (SOMs; Kohonen, 2001), which are a type of unsupervised artificial neural network used to perform data clustering, data-dimensionality reduction and data mining in a wide variety of application domains (Demartines and Herault, 1997; Gevrey et al., 2006; Zurita-Milla et al., 2013; Augustijn and Zurita-Milla, 2014).

In atmospheric sciences, SOMs have mostly been used to perform some kind of classification. For instance, they have been used to detect changes in wind trends whilst separating the contributions from ozone depletion and green house gas increases (Lee and Feldstein, 2013), to study El Niño Southern Oscillation-induced variation in tropical convection (Sakai and Iseri, 2010), to perform a climatological analysis of Northern Atlantic mean sea level pressure (Reusch et al., 2007), to relate increases in predicted precipitation in Greenland to changes in synoptic weather patterns (Schuenemann and Cassano, 2010) and to classify ozone profiles obtained with balloon sondes at two tropical sites (Jensen et al., 2012). However, to the best of our knowledge, SOMs have not been used for the application considered here despite the fact that they are likely more robust and effective than traditional methods. This is supported by Hsieh (2004) who compared nonlinear methods (including SOMs and other neural networks) and more traditional methods such as canonical correlation analysis, principal component analysis (PCA), rotated PCA, single spectrum analysis and Fourier spectrum analysis and showed that traditional methods may be limited in their capacity to capture geophysical patterns properly, especially when the data are no longer in the linear domain. Thus, SOMs might be better suited to point out weaknesses in retrieval algorithms caused by non-linear effects (e.g. abrupt changes caused by sensor degradation during the satellite's lifetime).

The remainder of this paper is organised as follows: Sect. 2 introduces the two data sets used to illustrate this study and explains the five steps of the approach. Section 3 provides details on how the approach was applied to the SCIAMACHY limb versus lidar ozone profile differences for each step and discusses the results. Section 4 presents our conclusions.

2 Data and methods

To illustrate the proposed SOM-based approach to intercompare data, we will use ozone profiles derived from SCIAMACHY limb measurements as well as from ground-based lidar stations. Sections 2.1 and 2.2 describe these two data sets and Sect. 2.3 will detail the five steps of the approach.

2.1 SCIAMACHY version 5.02 ozone profile data

SCIAMACHY stands for SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY. This instrument was launched into space on board ENVISAT, which was operational between March 2002 and April 2012, with first SCIAMACHY data from August 2002. SCIAMACHY is a passive remote sensing spectrometer observing backscattered, reflected, transmitted or emitted radiation from the Earth's surface and atmosphere, in the wavelength range between 240 and 2380 nm and in three measurement modes: occultation, nadir and limb geometry (Burrows et al., 1995; Bovensmann et al., 1999). In limb viewing mode, scans are made in steps of 3.3 km from (close to) the surface to an altitude of 92 km. The vertical resolution of the retrieved ozone profile product is typically ranging between 3 and 4 km. Here SCIAMACHY ozone number density data are extracted from the ozone profile product of the operational algorithm (level 2 version 5.02). The data retrieved in this version are most useful for altitudes between about 15 and 40 km because there is a reduced sensitivity to ozone above 40 km and below 20 km, leading to substantially increased retrieval errors at those altitudes (European Space Agency, 2011, 2013). The data are accompanied by quality flags indicating the validity and quality of the retrieved product (European Space Agency, 2013). Initial validation results for version 5.01 (for ozone profiles equivalent to version 5.02) showed a positive bias in the tropics, especially below 20 km, a good agreement (within 5 %) in the mid-latitudes and a variable bias was observed for the polar regions, with larger deviations above 35 km (European Space Agency, 2011; 2013).

2.2 Ground-based NDACC lidar data

In this study we have used ozone profiles obtained by ground-based lidars that are part of the Network for the Detection of Atmospheric Composition Change (NDACC; http://www.ndacc.org; Kurylo and Solomon, 1990). To be-

come associated with NDACC, it is obligatory to have a good description of the data quality through intercomparison of at least the retrieval software, followed by intercomparison with other instruments. The latter can be done with other instruments such as sondes, or with the NDACC travelling standard, the NASA GSFC (Goddard Space Flight Center) lidar. Ozone profiles are retrieved using differential absorption in sets of two wavelengths in the ultraviolet domain, where ozone is in each set only strongly absorbed at one of the two wavelengths. The difference of the slope of the logarithm of the retrieved lidar signals as a function of altitude is used to determine the ozone number density profile. The retrieval is as such self-calibrating. All of the lidars used operate under night-time conditions. The altitude range with the highest data quality is mostly ranging between 20 and 35 km (Keckhut et al., 2004), depending on laser power, optics and local atmospheric conditions. As the signal-to-noise ratio decreases with altitude, the retrieval error correspondingly increases, which can be partially compensated by degrading the vertical resolution. Typically the vertical resolution increases from several hundreds of metres in the lower stratosphere to several kilometres in the upper stratosphere (Godin et al., 1999).

In this study we used the following seven lidar sites (see Table 1 for their coordinates): Alomar (Norway), Hohenpeißenberg (Germany), Observatoire Haute Provence (France), Table Mountain (California), Mauna Loa (Hawaii), Lauder (New Zealand) and Dumont d'Urville (Antarctica). Four of the sites are located in the mid-latitudes, two in the polar regions and one in the tropics.

2.3 Methodology

The flowchart in Fig. 1 summarises the SOM-based approach as a series of five steps. In the first step, the data are prepared as input for the neural network. This involves the data selection (quality and collocation criteria), a calculation of the differences between the data sets and includes a data normalisation to set the variance to unity.

The normalised differences are used to train the self-organising map in the second step. The result of the training is a self-organised map where each neuron now has a normalised difference for each altitude that is similar to, but distinct from its neighbours, and is representative of one or multiple input vectors, but not necessarily identical to it (i.e. the vector of normalised differences is likely a weighted average of multiple input vectors) as also the neighbourhood affects the values assigned to one neuron. This set of representative normalised difference vectors is called the codebook vectors, which is a three-dimensional matrix (composed by the two dimensions of the SOM together with the altitude vector). For each altitude we can visualise these codebook vectors as a map, which will be called a component plane. In addition, we can also derive which neuron has the most similar normalised differences as the input data by finding the minimum

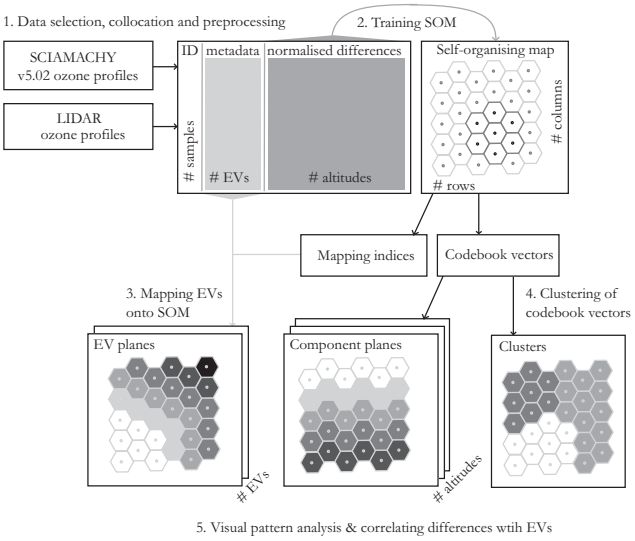

Figure 1. Flowchart of the proposed five-step methodology to explore origins of differences between ozone profiles. The self-organising map (SOM) is trained using a set of normalised differences (number of samples × number of altitudes). The SOM consists of a grid formed by a number of rows by columns that is defined by the user. In this example, hexagonal-shaped neurons are used resulting in six direct neighbours instead of four direct neighbours in a regular rectangular grid (as indicated with the darker grey hexagons in the upper right corner of the figure; the marked part corresponds to a radius of one neuron around the central neuron). The output of the training are the organised differences, called codebook vectors, which can be shown as a map (called component plane) for each altitude bin (variation in the relative differences at one altitude visualised by different grey-tones). Each data sample with associated explanatory variables (EVs) can be linked to a neuron on the SOM by the mapping indices. Using these mapping indices, the EVs can be projected onto the SOM (gradient in EV visualised by different grey-shades). The codebook vectors can be clustered to study sub-groups (here, three clusters are created and shown with different colours).

Euclidean distance between a data vector and codebook vectors, so it is known for each input data vector to which neuron it maps. This information is called the mapping index.

In the third step, the mapping indices are used to create maps of each explanatory variable (EV) with the same dimensions as the SOM. When multiple input vectors (IDs) map to the same neuron, it is necessary to calculate a mode, mean or median (depending on the type of variable) of the EV values of those IDs to associate to that neuron.

The codebook vectors can be clustered to help identify patterns that are present over the entire range of altitudes in the fourth step. Such a clustering is exemplified by the three colours in cluster block in Fig. 1.

The fifth and final part of the analysis is to study the relations between the component planes (the codebook vectors) and the explanatory variables, both on a global scale (entire data set) and on a more detailed (local) level inside the clus-

Table 1. Locations of the used lidar stations, with number of collocations per site used for the SOM analysis.

Site name	Latitude	Longitude	# of collocations
Alomar	69.3° N	16.0° E	423
Dumont d'Urville	66.6° S	140.0° E	73
Hohenpeißenberg	47.8° N	11.0° E	3800
Lauder	45.0° S	169.7° E	2239
Mauna Loa	19.5° N	155.6° W	4881
Observatoire Haute Provence	43.9° N	5.7° E	791
Table Mountain	34.4° N	117.7° W	1539

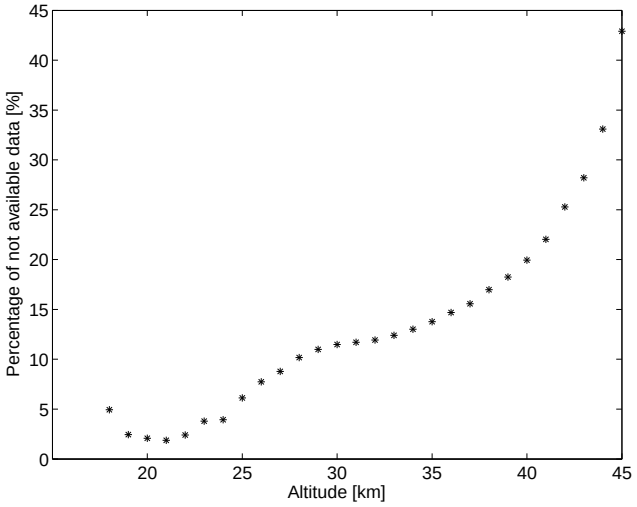

Figure 2. Fraction of not available difference data as a function of altitude.

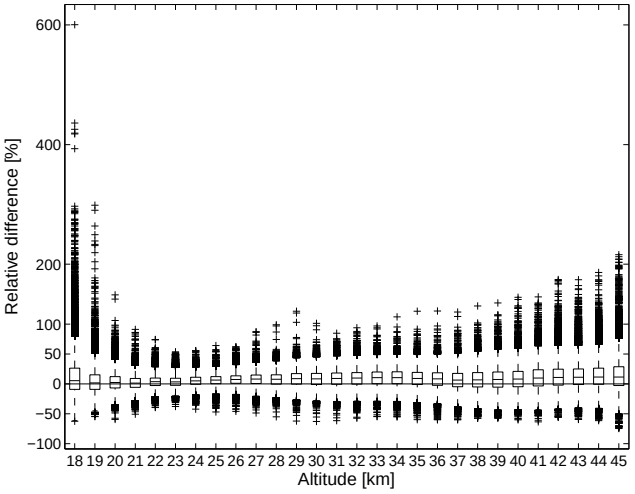

Figure 3. Box plot of the relative differences at a given altitude. The lower and upper boundaries of the boxes indicate the lower and upper quartiles. The lines between these boundaries correspond to the median. The dashed lines extending from the boxes show the range 1.5 times the interquartile range from the ends of each box. Outlier values outside this range are indicated with a +. The horizontal grey line indicates 0 % difference.

ters. This is done using visual inspection of the patterns in the codebook vectors and EVs and by calculating their Pearson linear correlations.

3 Practical implementation and discussion of results

To illustrate the analytical methods described in the previous section, in this section we present a detailed example following the five steps. Note that the third and fourth step can be executed in parallel (i.e. their relative order is arbitrary).

3.1 Data selection, collocation and preprocessing

We selected all SCIAMACHY and lidar ozone number density data from the period 2002–2012 having a reported error of 30 % or less and having valid processing flags. Collocations of SCIAMACHY and ground-based lidar ozone profiles were sought within 20 h and 800 km and the number of collocations with each site is listed in Table 1. The profiles are interpolated to a common altitude grid with a 1 km resolution using a nearly linear spline, followed by the calculation of the relative differences with respect to the lidar as follows: $\frac{\text{SCIAMACHY-lidar}}{\text{lidar}} \times 100\,\%$. The resulting data set consists of over 25 000 difference profiles between the collocated pairs, together with metadata (i.e. EVs) providing information on the observation characteristics. These differences are assumed to be to the largest extent attributable to the satellite retrieval as the lidar data are of high and known quality (Sect. 2.2). Here we further filtered the data to remove partial profiles; that is, where not for all altitudes between 18 and 45 km data were available for both the lidar and SCIAMACHY observations (see Fig. 2 where the fraction of not available data is indicated for each altitude).

This filtered data set consists of 13 746 difference profiles (with the matching metadata), which corresponds to 54 % of the input data having information for all selected altitudes. The histograms of the differences per altitude show close to Gaussian distributions, except for the lowest altitudes where the distribution is somewhat skewed, which is also visualised in the box plots shown in Fig. 3. Figure 4 shows the me-

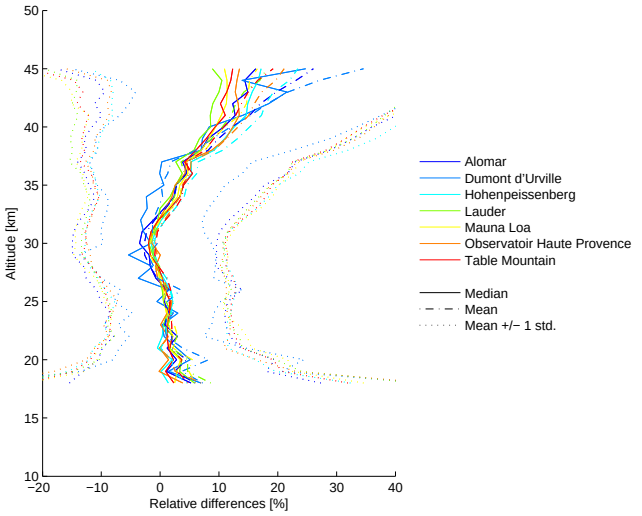

Figure 4. Relative differences between SCIAMACHY version 5.02 limb and lidar ozone profiles as a function of altitude colour-coded by collocated lidar site. The continuous lines indicate the median differences, the dash-dotted lines correspond to the mean differences and the dotted lines show the mean plus or minus one standard deviation of the relative differences.

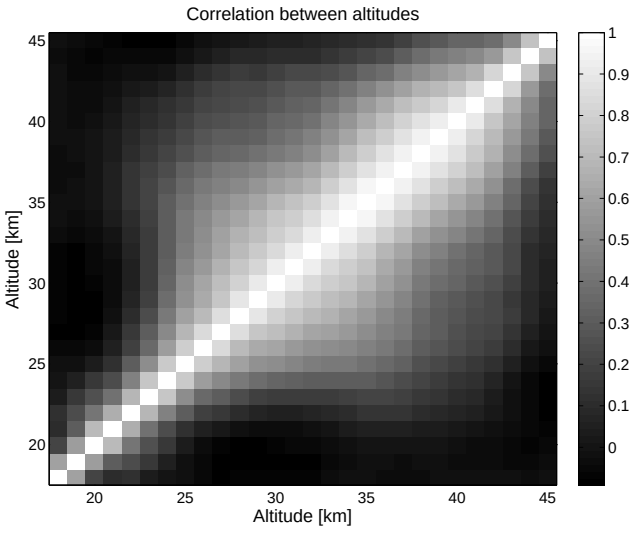

Figure 5. Correlation of the normalised relative differences between altitudes.

dian and mean (± one standard deviation) of the relative differences versus altitude colour coded by the collocated lidar site. The overall agreement between stations is very good at the altitude range where the lidar data are of highest quality. The most outstanding site is Dumont d'Urville, which besides being at an extreme location also has a limited number of collocations in our data selection. The top part (above 40 km) shows the largest divergence and increasing standard deviations which can be attributed to the increasing contribution of the a priori in the SCIAMACHY data.

The differences are normalised to set the variance to unity. As the transformation is linear, the distribution shapes are preserved. Figure 5 shows the Pearson linear correlation of the normalised relative differences between the 28 altitude bins. The correlation ranges between −0.09 and 0.95 (off-diagonal). It can be seen that differences at low altitudes are hardly correlated to those at higher altitudes and that at higher altitudes similar differences are found over a larger range of nearby altitudes.

3.2 Training of the SOM

The normalised data were used to train a SOM. Here we used the SOM toolbox for MATLAB version 2.0 beta by Alhoniemi, Himberg, Parhankangas and Vesanto available at http://www.cis.hut.fi/projects/somtoolbox. The self-organising map was set up as a lattice grid of 46 by 75 hexagonal neurons.

The dimensions were chosen to theoretically allow an average of four input vectors to map onto a single neuron, which was chosen as a trade-off between complexity and over-simplification. Different ratios determine the level of detail that can be studied, but the principles remain the same. In the case of a greatly extended geographical input space, one could additionally go for a high-resolution representation with more neurons than input vectors (Skupin and Esperbé, 2011).

The training was done in two phases. The initial phase consisted of 200 iterations where a rough training was carried out with an initial neighbourhood covering a radius of 10 neurons which gradually decreased to cover a radius of 2.5 neurons at the end of this phase. The second, fine-tuning phase was then run for 400 iterations with a neighbourhood covering a radius of 2.5 neurons gradually decreasing to a radius of a single neuron at the end of the training. In both cases we have used the batch training algorithm.

Most of the neurons get organised with inputs from one (37 %) to two (38 %) sites, about 18 % with inputs from three sites and very few (less than 5 % in total) can be related to four or five sites. Assignation of six or all seven sites to the same neuron does not occur. More than 61 % of the neurons are mapped to by multiple sites. Overall, this indicates that the relative differences between SCIAMACHY and the lidar ozone profiles appear to be (indirectly) location dependent to a limited extent. No input data get mapped onto 81 neurons (∼ 2 %), which indicates that some difference values in the SOM space do not occur.

One large advantage of the method used here is that number of collocating profiles per station is not required to be the same. In fact, this method actually minimises the impact of having an uneven distribution of number of collocations per site. If, for instance, a very large number of observations for a particular site result in very similar differences, these will map to a small set of contiguous neurons whereas in tradi-

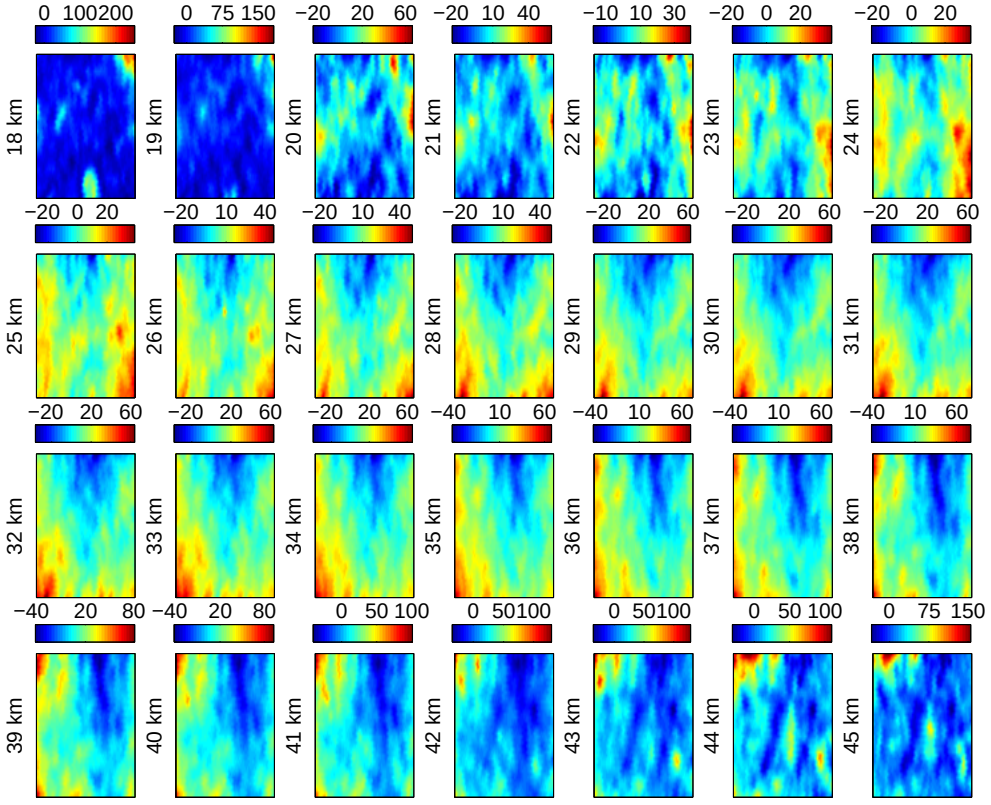

Figure 6. De-normalised codebook vectors for the 28 altitudes ascending from left to right row wise. Note that the range of colours is optimised for each panel; the associated relative differences (%) are indicated by the colour legend on top of each panel. The component planes are interpolated to avoid empty (non-used) neurons and bridge the gap in the input space. The maps are stretched in the vertical direction for a better visualisation.

tional analyses they would heavily influence the outcome of the correlation.

3.2.1 Component planes

Figure 6 presents the component planes for the 28 altitudes (18 km in the upper left corner, 45 km in the lower right corner). The codebook vectors have been de-normalised, so that the units correspond to the original relative differences. We can observe that at the lowest altitudes (18 and 19 km) most neurons have values for the relative differences that are close to zero, but some spots with higher deviations stand out like the upper right corner with the most extreme outliers and multiple regions (cyan-coloured) with a similar positive deviation. With ascending altitude, we can see that although at low altitudes these differences were similar, higher up they are clearly distinct. For instance, the blob with positive differences at the bottom of the 18 km panel is associated with negative deviations around 20–21 km and higher up with relatively small deviations from zero whereas the blob at the upper right corner remains a positive bias for many kilometres.

The organisation of the differences has also led to a small increase of the correlation between near altitudes as less representative samples get to play a smaller role. Overall patterns (as visible in Fig. 5 and its discussion) are nevertheless preserved.

3.3 Mapping the explanatory variable (EV) planes

Using the mapping indices and the IDs of each data sample, we can link the explanatory variables to the neuron where the corresponding set of relative differences for the 28 altitudes mapped. In this way, the information is summarised and organised following the spatial structure defined during the training of the SOM using the relative differences. This allows us to visually identify patterns and the relative importance of the selected EVs. Here we have considered 11 variables which are also often used in traditional validation studies (see Table 2).

When more than one input data sample maps onto the same neuron, a representative EV value was calculated considering the data type of the EV. For the scan direction, the location and the day of the year, we used the mode function. For the rest of the EVs, the mean function was used.

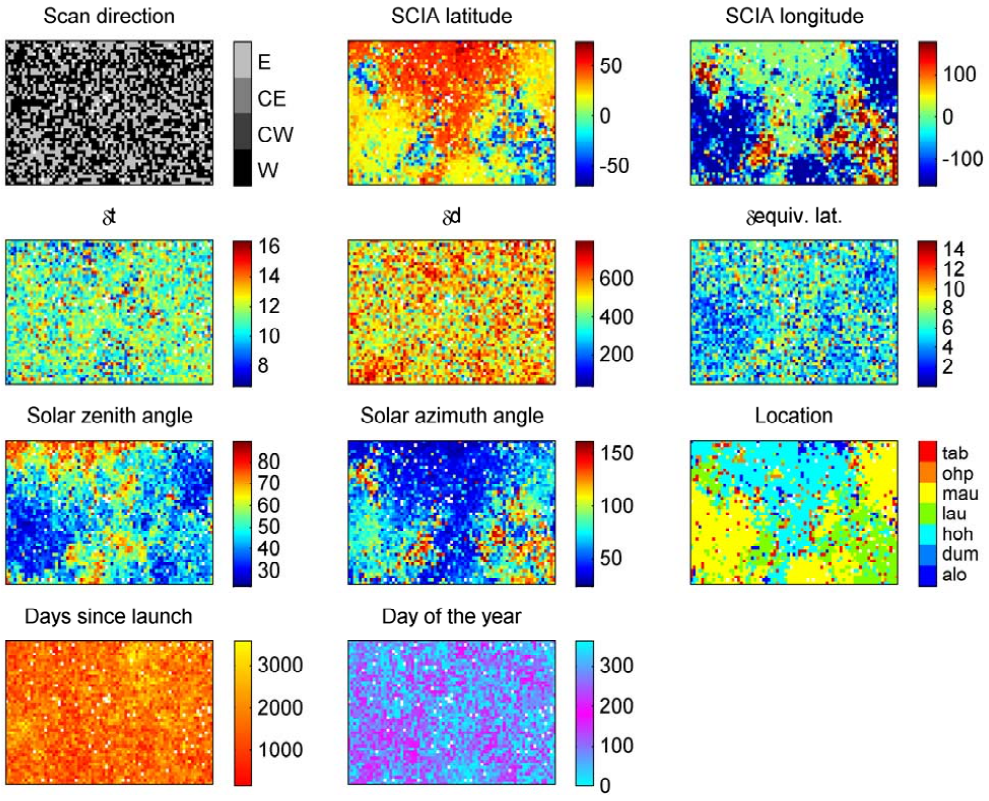

Figure 7. Explanatory variables mapped on the SOM. From left to right: top row: SCIAMACHY scan direction, SCIAMACHY latitude, SCIAMACHY longitude; second row: difference in time between collocations, difference in distance between collocations, difference in equivalent latitude between collocations; third row: solar zenith angle during SCIAMACHY observation, solar azimuth angle during SCIAMACHY observation, lidar station name; lower row: days since launch of ENVISAT, day of the year. White pixels indicate neurons that are not used.

Table 2. Used explanatory variables with corresponding range of values.

Variable name	Minimum value	Maximum value
SCIAMACHY scan direction (4 profiles retrieved E-W)	West	East
Latitude of the SCIAMACHY observation	74° S	76° N
Longitude of the SCIAMACHY observation	163° W	180° E
Difference in time between collocations	−17.8 h	+18.5 h
Difference in horizontal direction between collocations	4 km	800 km
Difference in equivalent latitude between collocations	0°	15°
Solar zenith angle during the SCIAMACHY observation	21.9°	89.8°
Solar azimuth angle during the SCIAMACHY observation	17.3°	164.3°
Station's name of the collocated lidar ozone profile	–	–
Number of days since the launch of ENVISAT	154 d	3583 d
Day of the year	0	364

The difference in time between the lidar and SCIAMACHY observation was taken as an absolute difference, disregarding whether the lidar or SCIAMACHY observation was acquired first. Using other statistics such as the median instead of the mean, did not affect the patterns much, indicating that the organisation is consistent. Various of the EVs are correlated to some extent, which can be expected because of the nature of these variables and the limited number of lidar sites available here. Nevertheless, most of the ranges of variation in the EVs is observed at more than one site and we deem the covered variation sufficient to study the dependencies between the ozone profile differences and the EVs.

Figure 7 shows the 11 EVs mapped onto the SOM. The white dots represent empty neurons (no data mapped onto

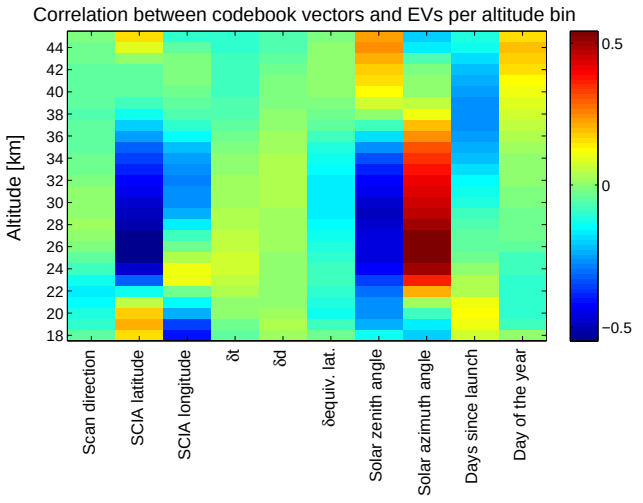

Figure 8. Pearson linear correlation between the codebook vectors at a given altitude and the mapped explanatory variables (EVs). EVs from left to right: scan direction, latitude of the SCIAMACHY observation, longitude of the SCIAMACHY observation, difference in time between collocations, difference in distance between collocations, difference in equivalent latitude between collocations, solar zenith angle during SCIAMACHY observation, solar azimuth angle during SCIAMACHY observation, days since the launch of ENVISAT and the day of the year. See also scatter plots of various EVs versus the codebook vectors at selected altitudes in the Supplement.

them). Some variables like the scan direction and difference in time or distance between the observations appear to lack any organisation. In other words, they exhibit a random distribution in the SOM space. Other EVs, like the latitude of the SCIAMACHY observation, show a clear organisation and thus are significant to explain the organisation of the relative differences. We can also see that some of the EV patterns are similar (for example location and longitude). This evidences a relationship between the different EVs. A third group of EVs is composed of those variables that appear to be quite randomly distributed except for some spots where they get grouped consistently. Days since launch and difference in equivalent latitude belong to this third EV group, which points at secondary dependences or perhaps indicates a behaviour that is local rather than global.

The downward pointing triangular area that can be seen in the solar azimuth angle (low angles; dark blue) and in the coordinates of the SCIAMACHY observation (latitude of about 50°, red; mean longitude of around 0°, green) mapped onto the SOM appears to be linked with a similarly shaped area in the component planes of altitudes of 30 ± 5 km showing a small to negligible underestimation of the ozone profiles by SCIAMACHY.

3.4 Correlation hunting

The patterns in the codebook vectors and the mapped EVs can be directly compared by calculating the Pearson correlation coefficient between each EV and the codebook vector for a given altitude. This provides a convenient way to visualise the linear dependence of the differences data set on a given variable at a specific altitude. However, it is not the same as using the original input data to calculate such correlations, as the data have been summarised and approximated by the codebook vectors when they were organised. The mapped EVs represent a synopsis of the original data in a similar manner.

Figure 8 presents the correlations between the component planes and the EVs for the global data set. We can see that at the lowest altitudes studied here, the relative differences seem to be most dependent on the longitude of the SCIAMACHY observations. Continuing upwards, the most dominant, but relatively weak, factor becomes the solar zenith angle, followed by the latitude of the SCIAMACHY observation at the ozone maximum. We can see that the solar azimuth angle is strongly coupled to the latitude of the observation and the solar zenith angle to a lesser extent as well. Then above 36 km, the differences are in some part organised according to the age of the sensor (days since launch). At the highest altitudes, we see again the solar zenith angle having the largest correlation with the organised differences, but also some seasonal effect through the day of the year. Scatter plots of selected EVs versus the codebook vectors of the relevant altitudes are provided in the Supplement. The data selection procedure is also shown to be quite adequate, as no dependence is observed on the difference in time or space between the lidar and SCIAMACHY observations. There is however a small contribution to the distribution of differences by the difference in equivalent latitude, mostly around the ozone maximum and around 20 km. The effect of the scan angle on the distribution of the differences on a global basis seems to be marginal and is only somewhat influencing around 22 km.

To study whether the observed correlations can be partially explained by a correlation of the EV with latitude or latitude and longitude, we performed a partial correlation analysis for the other EVs where the correlation between the EV and the codebook vector with latitude (and longitude) is taken into account following

$$\rho_{\mathrm{EC,L}} = \frac{\rho_{\mathrm{EC}} - \rho_{\mathrm{EL}}\rho_{\mathrm{CL}}}{\sqrt{\left(1-\rho_{\mathrm{EL}}^2\right)\left(1-\rho_{\mathrm{CL}}^2\right)}}, \quad (1)$$

with E equal to the EV of interest, C the matrix of codebook vectors, L the latitude (+longitude) and ρ the Pearson linear correlation coefficient.

Figure 9 shows the results of the partial correlation analysis taking the correlation with both latitude and longitude into account. At the lowest altitude, the most important remaining

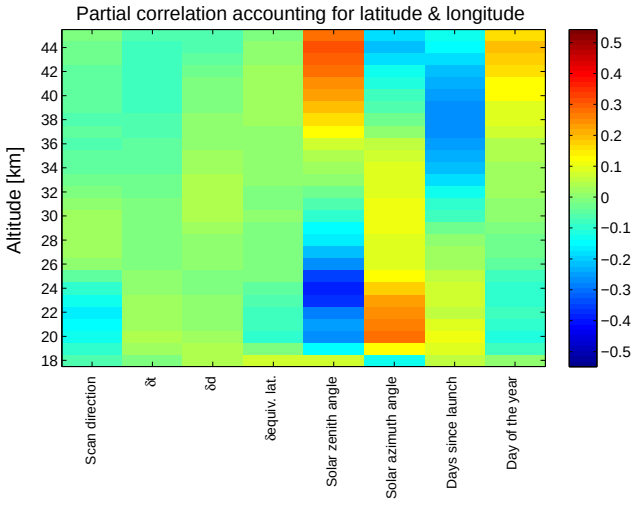

Figure 9. Partial Pearson correlations between the codebook vectors and EVs (from left to right: SCIAMACHY scan direction, difference between collocations in time, distance and equivalent latitude, solar zenith and azimuth angles, days since the launch of ENVISAT and day of the year) as a function of altitude after removing correlations with latitude and longitude. The colour bar is scaled to the same range as Fig. 8.

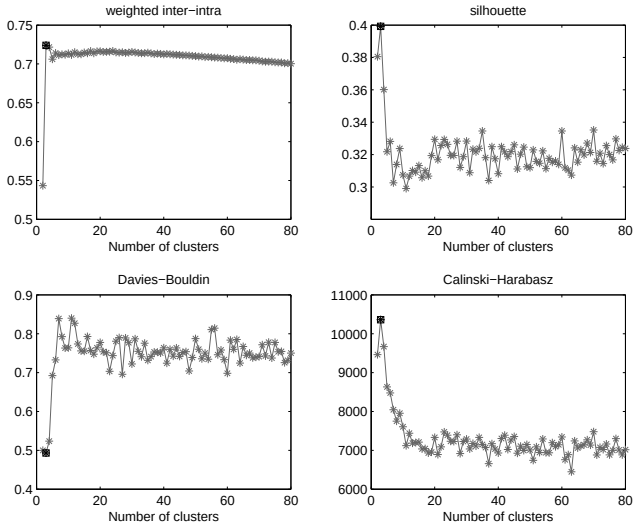

Figure 10. Internal validity indices for k-means clustering of the codebook vectors into 2 to 80 clusters. Left to right and top downward: weighted inter-intra index, silhouette index, Davies–Bouldin index and the Caliński–Harabasz index. The optimum number of classes specified by a given index is indicated with a black square and data point.

factor is the solar azimuth angle, followed by the solar zenith angle when going upwards. Between 26 and 36 km, most of the variation seems to be captured by the combination of latitude and longitude as the residual partial correlations are low. However, (partial) correlations cannot point out the mechanisms that are actually driving the observed differences in ozone profiles (i.e. the causal factors). Above 36 km, some residual variation can be attributed to the time since launch and then again to the solar zenith angle. To a lesser extent, also the day of the year and the solar azimuth angle remain influential. When considering correlation with latitude alone (not shown), we observe that only the correlation for solar azimuth angle is greatly linked to latitude. For longitude, we observe that the partial correlation has reduced the Pearson correlation values below 25 km (except at 22 km) and above 43 km, but correlation has become stronger at all altitudes in between. Overall, the partial correlation shows a stronger link with longitude. For solar zenith angles, we see almost no changes.

Naturally, there will be more factors involved in the distribution of differences between the lidar and SCIAMACHY ozone profiles, for instance influences from the level 1 processing (calibration errors) or the sensor's sensitivity. Our intention here is to demonstrate the possibilities of the proposed approach and we have thus limited the explanatory variables to those commonly selected in validation studies. We note that when additional information becomes available during the course of a study, the approach allows us to very simply and quickly extend the analysis with those new variables. Studying the physical basis of the retrieval of SCIAMACHY limb ozone profiles is beyond the scope of this paper.

3.5 Clustering

We have clustered the codebook vectors using simple k-means. Different numbers of clusters were tested for consistency and the experiment was repeated 100 times to see the stability of the obtained solutions as the clustering may be sensitive to the initialisation/seeding values and get trapped in a local optimum (Tzortzis and Likas, 2014). Various algorithms exist that try to optimise the number of clusters, based on the principle that similarity is indicated by the inter- and/or intra-cluster distances in the data space (e.g. Davies and Bouldin, 1979; Tibshirani et al., 2001; Caliński and Harabasz, 1974). However, no clear optimum might be found if the clusters present gradual transitions. A further limitation is that patterns visible to the eye at a single altitude may not be identified by the clustering algorithm run on the entire set of codebook vectors as one altitude has a relatively low weight on the total data set.

Four indices were chosen to examine the clustering efficiency: the Davies–Bouldin index (considers the ratio of the intra-cluster scatter to the inter-cluster separation), the weighted inter-intra index (the ratio of weighted average inter-cluster to weighted average intra-cluster similarity), the silhouette index (measure of how close each point in one cluster is to points in the neighbouring clusters) and the Caliński–Harabasz index (considers the ratio of the inter-cluster variance to the intra-cluster variance). The Davies–

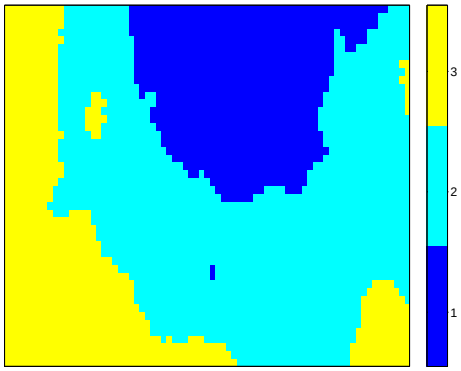

Figure 11. *k*-means classification of the codebook vectors into three classes.

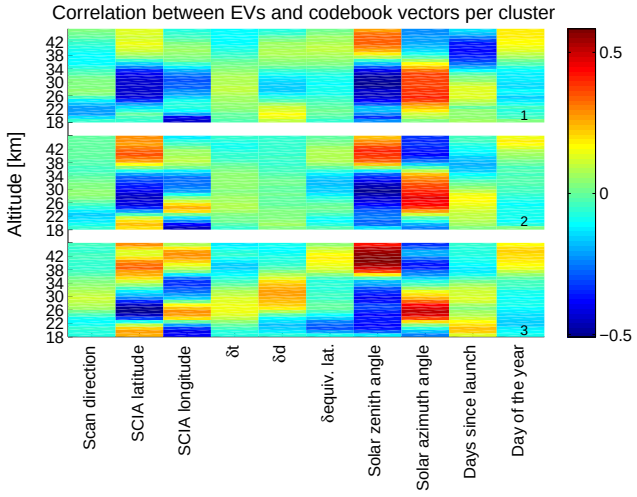

Figure 12. Pearson linear correlation between the codebook vectors at a given altitude and the mapped explanatory variables (EVs) for the three clusters shown in Fig. 9 (first cluster on top, etc.) as indicated in the lower right corner of each subplot. EVs from left to right: scan direction, latitude of the SCIAMACHY observation, longitude of the SCIAMACHY observation, difference in time between collocations, difference in distance between collocations, difference in equivalent latitude between collocations, solar zenith angle during SCIAMACHY observation, solar azimuth angle during SCIAMACHY observation, days since the launch of ENVISAT and the day of the year.

Bouldin index has to be minimised whereas the other three indices have to be maximised. Figure 10 shows the values for the four indices when running a *k*-means classification of the codebook vectors for two to 80 clusters. All four indices indicate that the optimal number of clusters is equal to three.

Figure 11 shows the resulting clustering obtained in 80 % of the runs. Two small "islands" of other clusters can be seen inside the second cluster. This could be due to the clustering not being totally successful (for instance, due to gradual transitions of the data) or an imperfect organisation by the SOM.

Alternatively, one could argue that visually more than three clusters can be identified given the patterns in the component planes where clearly small groups can be seen at certain altitudes. The choice is therefore depending on the level of detail required by the user. For the purpose of illustration and comparison between clusters, visualisation of three clusters is assumed to be sufficiently adequate besides this choice being supported by the cluster validity indices.

Additionally, clustering could also be done based on an EV when the differences have been shown to be dependent on this variable. Such a sub-selection is then to be defined by the user.

3.5.1 Cluster-wise correlation hunting

Section 3.4 has presented the correlations for the full set of explanatory variables and the SOM's component planes, providing a global overview. More complex relations might be obtained when examining the relations inside clusters. It is possible that some parts of the data set respond to different variables (local relations) or behave in an opposite manner, which then do not show up as a significant correlation in the global analysis.

We have repeated the correlation analysis for the three clusters created from the codebook vectors in the previous section.

Figure 12 shows the linear correlations between the codebook vectors and the mapped EVs for the three clusters. We can see that some details are very distinct for the different clusters. For instance, the dependence on latitude at higher altitudes appears to be much lower for the first cluster in comparison to the other two clusters. This cluster mostly contains data originating from the northern mid-latitudes, yet this limited latitudinal coverage does not affect the correlation for altitudes between 24–34 km substantially and another process must be responsible for this. In contrast, a substantial correlation (-0.37) with the days since launch between 36 and 42 km has appeared in cluster 1 besides the previously detected correlation with the solar zenith angle. Also the dependence on the scan direction is stronger for the different clusters than for the global data set. For the third cluster some correlation with the distance between the SCIAMACHY and lidar observations has appeared for the middle part of the selected altitude range, coincident with a stronger correlation with longitude and with the scan angle for those altitudes. This should be studied in more detail. Another observation is that for the full altitude range, the strongest dependence on latitude (highest negative correlation) and solar zenith angle (highest positive correlation) is found for the third cluster, which also contains the largest variation (standard deviation) in latitudes covered by the input data. The solar zenith angle has a greater negative correlation for the first cluster. Observations such as these made here should be of interest for the teams working on the retrieval algorithms, as they can focus

on studying why these parameters/variables have such an impact on the observed differences between SCIAMACHY and the lidar ozone profiles and reveal whether there is a physical basis for the results found in this exploratory study.

4 Summary and conclusions

In this paper we have presented an alternative and novel approach to intercompare data sets and explore (dis)similarity patterns and their possible causes. The approach is based on the use of self-organising maps (SOMs) and was applied to atmospheric ozone profiles (satellite and ground-based). More precisely, the proposed approach was illustrated using SCIAMACHY limb ozone profile data (level 2 version 5.02 of the operational product) and ground-based lidar ozone profiles around seven observation sites.

Following profile collocation and data quality filtering, relative differences between the two instruments were calculated for altitudes between 18 to 45 km, and subsequently normalised using the variance (preserving the distribution). A SOM was batch-trained in two phases using these normalised relative differences.

After that, the role of 11 selected explanatory variables (EVs, related to location, data collocation criteria and observational characteristics) was studied by mapping these onto the trained SOM. For this, we relied on the ID that links each set of ozone profile differences to the explanatory variables. Through visual inspection of the patterns formed on the two-dimensional SOM and through correlation analysis, relations between the self-organised differences and the various of explanatory variables became apparent, directly linking the differences between the SCIAMACHY and lidar profiles with these EVs without having to a priori specify certain conditions or ranges of values for the explanatory variables (current common practice in data intercomparison exercises). Some of the EVs are to a certain extent correlated because of the nature of these observations (e.g. latitude and solar azimuth angle) or through introduction by the data selection procedure, but the variation covered was considered sufficient to attempt to identify dependencies between the observed ozone profile differences and the different EVs.

At the lowest altitudes studied here, the largest influencing factor determining the patterns of relative differences seem to be the longitude of the SCIAMACHY observations. Higher up, the most dominant, but relatively weak, factor become the solar zenith angle, followed by the latitude of the SCIAMACHY observation at the ozone maximum. The solar azimuth angle is strongly coupled to the latitude of the observation (as confirmed by partial correlation analysis accounting for latitude) and the solar zenith angle to a lesser extent as well. Above 36 km, the differences are in some part organised according to the age of the sensor (days since launch). At the highest altitudes, the solar zenith angle again has the largest correlation with the organised differences, but also some seasonal effect appears to play a role visualised by the day of the year. The data selection procedure is also shown to be quite adequate, as no dependence is observed on the difference in time or space between the lidar and SCIAMACHY observations. There is however a small contribution to the distribution of differences in the ozone profiles by the difference in equivalent latitude between the two observations, mostly around the ozone maximum and around 20 km. The effect of the scan angle on the distribution of the differences on a global basis seems to be marginal and is only somewhat influencing around 22 km. A partial correlation analysis adjusting for latitude and longitude yielded residual correlations for solar azimuth and zenith angles, the day of the year and the SCIAMACHY's sensor age. At the lowest altitude, the most influencing factor is then the solar azimuth angle, followed by the solar zenith angle. No dominant factor is found between 26 and 36 km, and as with the standard correlation analysis, above 36 km the sensor age and then solar zenith angle have the highest correlation. However, the correlation analysis cannot point out which factor, or combination of factors, drives the observed differences between the ozone profiles as most of the factors are inter-related and identification of the underlying mechanisms should be studied in the retrieval algorithms.

Further details were obtained by clustering the SOM component planes (i.e. the values of the self-organised differences at multiple heights) into three clusters and investigating the differences between the clusters in terms of the explanatory variables. Although the general patterns are similar, some of the details are very distinct for the different clusters. For example, in the first cluster the dependence on latitude at higher altitudes appears to be much lower in comparison to the other clusters. The first cluster mostly contains data originating from the northern mid-latitudes, yet this limited latitudinal coverage does not substantially affect the correlation for altitudes between 24–34 km and it is postulated that another process must be responsible for this. In contrast, a substantial negative correlation with the days since launch between 36 and 42 km appears for this cluster in addition to the already detected positive correlation with solar zenith angle. Also the dependence on the scan direction is stronger for the different clusters than for the global data set. For the third cluster some correlation with the distance between the SCIAMACHY and lidar observations shows up for the middle part of the selected altitude range, coincident with a stronger correlation with longitude and with the scan angle for those altitudes, which should be further studied.

Summarising, the SOM-based approach has shown the potential to study relations between observed differences between two data sets and possible underlying factors without making prior assumptions on which factors are of interest whilst reducing the introduction of biases due to an uneven distribution of collocations per site. It is simple to add EVs to the analysis and no a priori division into ranges of values for the variables is required. The level of detail can be opti-

mised by adjusting the SOM size and, by looking at clusters inside the SOM, local relations may be studied. The proposed approach is thus offering a fresh and unbiased look at differences between data sets and is very useful to point out where further focus should be laid to investigate the origins of the differences and to enhance the underlying algorithms and/or models.

The Supplement related to this article is available online at doi:10.5194/amt-8-1951-2015-supplement.

Acknowledgements. The authors would like to acknowledge financial support from the European Space Agency (ESA) through the VALID-2 project and the Netherlands Space Office (NSO) through the NL-SCIAvisie project led by Ilse Aben (Netherlands Institute for Space Research (SRON)). We would also like to thank the two anonymous reviewers for their feedback.

Edited by: M. Weber

References

Adams, C., Strong, K., Batchelor, R. L., Bernath, P. F., Brohede, S., Boone, C., Degenstein, D., Daffer, W. H., Drummond, J. R., Fogal, P. F., Farahani, E., Fayt, C., Fraser, A., Goutail, F., Hendrick, F., Kolonjari, F., Lindenmaier, R., Manney, G., McElroy, C. T., McLinden, C. A., Mendonca, J., Park, J.-H., Pavlovic, B., Pazmino, A., Roth, C., Savastiouk, V., Walker, K. A., Weaver, D., and Zhao, X.: Validation of ACE and OSIRIS ozone and NO_2 measurements using ground-based instruments at 80° N, Atmos. Meas. Tech., 5, 927–953, doi:10.5194/amt-5-927-2012, 2012.

Augustijn, P. W. M. and Zurita-Milla, R.: Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns, Int. J. Health Geogr., 12, 60, doi:10.1186/1476-072X-12-60, 2014.

Boersma, K. F., Jacob, D. J., Eskes, H. J., Pinder, R. W., Wang, J., and van der A, R. J.: Intercomparison of SCIAMACHY and OMI tropospheric NO_2 columns: observing the diurnal evolution of chemistry and emissions from space, J. Geophys. Res., 113, D16S26, doi:10.1029/2007JD008816, 2008.

Bovensmann, H., Burrows, J. P., Buchwitz, M., Frerick, J., Noël, S., Rozanov, V. V., Chance, K. V., and Goede, A. P. H.: SCIAMACHY: mission objectives and measurement modes, J. Atmos. Sci., 56, 127–150, doi:10.1175/1520-0469(1999)056<0127:SMOAMM>2.0.CO;2, 1999.

Burrows, J. P., Hölzle, E., Goede, A. P. H., Visser, H., and Fricke, W.: SCIAMACHY – Scanning Imaging Absorption Spectrometer for Atmospheric Chartography, Acta Astronaut., 35, 445–451, doi:10.1016/0094-5765(94)00278-T, 1995.

Caliński, T. and Harabasz, J.: A dendrite method for cluster analysis, Commun. Stat., 3, 1–27, doi:10.1080/03610927408827101, 1974.

Davies, D. L. and Bouldin, D. W.: A cluster separation measure, IEEE T. Pattern Anal., PAMI-1, 224–227, doi:10.1109/TPAMI.1979.4766909, 1979.

Demartines, P. and Herault, J.: Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets, IEEE T. Neural Networ., 8, 148–154, doi:10.1109/72.554199, 1997.

European Space Agency: Disclaimer for SCIAMACHY Level 2 data version SCIAMACHY/OL5.02 (ENVI-GSOP-EOGD-QD-11-0110), available at: https://earth.esa.int/documents/10174/24074/SCI_OL__2P_README.pdf (last access: 12 January 2014), 2011.

European Space Agency: Readme file for SCIAMACHY Level 2 version 5.02 products – Issue 1.2 (ENVI-GSOP-EOGD-QD-13-0118), available at: https://earth.esa.int/handbooks/availability/disclaimers/SCI_OL__2P_README.pdf (last access: 12 January 2014), 2013.

van Gijsel, J. A. E., Swart, D. P. J., Baray, J.-L., Bencherif, H., Claude, H., Fehr, T., Godin-Beekmann, S., Hansen, G. H., Keckhut, P., Leblanc, T., McDermid, I. S., Meijer, Y. J., Nakane, H., Quel, E. J., Stebel, K., Steinbrecht, W., Strawbridge, K. B., Tatarov, B. I., and Wolfram, E. A.: GOMOS ozone profile validation using ground-based and balloon sonde measurements, Atmos. Chem. Phys., 10, 10473-10488, doi:10.5194/acp-10-10473-2010, 2010.

Gevrey, M., Worner, S., Kasabov, N., Pitt, J., and Giraudel, J.-L.: Estimating risk of events using SOM models: a case study on invasive species establishment, Ecol. Model., 197, 361–372, doi:10.1016/j.ecolmodel.2006.03.032, 2006.

Godin, S., Carswell, A. I., Donovan, D. P., Claude, H., Steinbrecht, W., McDermid, I. S., McGee, T. J., Gross, M. R., Nakane, H., Swart, D. P. J., Bergwerff, H. B., Uchino, O., von der Gathen, P., and Neuber, R.: Ozone differential absorption lidar algorithm intercomparison, Appl. Optics, 38, 6225–6236, doi:10.1364/AO.38.006225, 1999.

Herman, J., Cede, A., Spinei, E., Mount, G., Tzortziou, M., and Abuhassan, N.: NO_2 column amounts from ground-based Pandora and MFDOAS spectrometers using the direct-sun DOAS technique: intercomparisons and application to OMI validation, J. Geophys. Res., 114, D13307, doi:10.1029/2009JD011848, 2009.

Hsieh, W. W.: Nonlinear multivariate and time series analysis by neural network methods, Rev. Geophys., 42, RG1003, doi:10.1029/2002RG000112, 2004.

Jensen, A. A., Thompson, A. M., and Schmidlin, F. J.: Classification of Ascension Island and Natal ozonesondes using self-organizing maps, J. Geophys. Res., 117, D04302, doi:10.1029/2011JD016573, 2012.

Keckhut, P., McDermid, I. S., Swart, D. P. J., McGee, T. J., Godin-Beekmann, S., Adriani, A., Barnes, J., Baray, J.-L., Bencherif, H., Claude, H., Fiocco, G., Hansen, G. H., Hauchecorne, A., Leblanc, T., Lee, C. H., Pal, S., Mégie, G., Nakane, H., Neuber, R., Steinbrecht, W., and Thayer, J.: Review of ozone and temperature lidar validations performed within the framework of the Network for the Detection of Stratospheric Change, J. Environ. Monitor., 6, 721–733, doi:10.1039/B404256E, 2004.

Kohonen, T.: Self-Organizing Maps, 3rd Edn., Springer-Verlag, New York, USA, 502 pp., 2001.

Kurylo, M. J. and Solomon, S.: Network for the detection of stratospheric change: a status and implementation report, Issued by NASA Upper Atmosphere Research Program and NOAA Climate and Global Change Program, Washington DC, 1990.

Lamsal, L. N., Martin, R. V., van Donkelaar, A., Celarier, E. A., Bucsela, E. J., Boersma, K. F., Dirksen, R., Luo, C., and Wang, Y.: Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: insight into the seasonal variation of nitrogen oxides at northern midlatitudes, J. Geophys. Res., 115, D05302, doi:10.1029/2009JD013351, 2010.

Lee, S. and Feldstein, S. B.: Detecting ozone- and greenhouse gas-driven wind trends with observational data, Science, 339, 563, doi:10.1126/science.1225154, 2013.

Nair, P. J., Godin-Beekmann, S., Froidevaux, L., Flynn, L. E., Zawodny, J. M., Russell III, J. M., Pazmiño, A., Ancellet, G., Steinbrecht, W., Claude, H., Leblanc, T., McDermid, S., van Gijsel, J. A. E., Johnson, B., Thomas, A., Hubert, D., Lambert, J.-C., Nakane, H., and Swart, D. P. J.: Relative drifts and stability of satellite and ground-based stratospheric ozone profiles at NDACC lidar stations, Atmos. Meas. Tech., 5, 1301–1318, doi:10.5194/amt-5-1301-2012, 2012.

Nazaryan, H., McCormick, M. P., and Russell III, J. M.: Comparative analysis of SBUV/2 and HALOE ozone profiles and trends, J. Geophys. Res., 112, D10304, doi:10.1029/2006JD007367, 2007.

Reusch, D. B., Alley, R. B., and Hewitson, B. C.: North Atlantic climate variability from a self-organizing map perspective, J. Geophys. Res., 112, D02104, doi:10.1029/2006JD007460, 2007.

Sakai, K., Kawamura, R., and Iseri, Y.: ENSO-induced tropical convection variability over the Indian and western Pacific oceans during the northern winter as revealed by a self-organizing map, J. Geophys. Res., 115, D19125, doi:10.1029/2010JD014415, 2010.

Schuenemann, K. C. and Cassan, J. J.: Changes in synoptic weather patterns and Greenland precipitation in the 20th and 21st centuries: 2. Analysis of 21st century atmospheric changes using self-organizing maps, J. Geophys. Res., 115, D05108, doi:10.1029/2009JD011706, 2010.

Skupin, A. and Esperbe, A.: An alternative map of the United States based on an n-dimensional model of geographic space, J. Visual. Lang. Comput., 22, 290–304, doi:10.1016/j.jvlc.2011.03.004, 2011.

Stiller, G. P., Kiefer, M., Eckert, E., von Clarmann, T., Kellmann, S., García-Comas, M., Funke, B., Leblanc, T., Fetzer, E., Froidevaux, L., Gomez, M., Hall, E., Hurst, D., Jordan, A., Kämpfer, N., Lambert, A., McDermid, I. S., McGee, T., Miloshevich, L., Nedoluha, G., Read, W., Schneider, M., Schwartz, M., Straub, C., Toon, G., Twigg, L. W., Walker, K., and Whiteman, D. N.: Validation of MIPAS IMK/IAA temperature, water vapor, and ozone profiles with MOHAVE-2009 campaign measurements, Atmos. Meas. Tech., 5, 289–320, doi:10.5194/amt-5-289-2012, 2012.

Takele Kenea, S., Mengistu Tsidu, G., Blumenstock, T., Hase, F., von Clarmann, T., and Stiller, G. P.: Retrieval and satellite intercomparison of O_3 measurements from ground-based FTIR Spectrometer at Equatorial Station: Addis Ababa, Ethiopia, Atmos. Meas. Tech., 6, 495–509, doi:10.5194/amt-6-495-2013, 2013.

Tibshirani, R., Walther, G., and Hastie, T.: Estimating the number of clusters in a dataset via the Gap statistic, J. R. Stat. Soc. B, 63, 411–423, doi:10.1111/1467-9868.00293, 2001.

Tzortzis, G. and Likas, A.: The MinMax *k*-means clustering algorithm, Pattern Recogn., 47,2505–2516, doi:10.1016/j.patcog.2014.01.015, 2014.

Wetzel, G., Oelhaf, H., Berthet, G., Bracher, A., Cornacchia, C., Feist, D. G., Fischer, H., Fix, A., Iarlori, M., Kleinert, A., Lengel, A., Milz, M., Mona, L., Müller, S. C., Ovarlez, J., Pappalardo, G., Piccolo, C., Raspollini, P., Renard, J.-B., Rizi, V., Rohs, S., Schiller, C., Stiller, G., Weber, M., and Zhang, G.: Validation of MIPAS-ENVISAT H_2O operational data collected between July 2002 and March 2004, Atmos. Chem. Phys., 13, 5791–5811, doi:10.5194/acp-13-5791-2013, 2013.

Zhang, L., Jacob, D. J., Liu, X., Logan, J. A., Chance, K., Eldering, A., and Bojkov, B. R.: Intercomparison methods for satellite measurements of atmospheric composition: application to tropospheric ozone from TES and OMI, Atmos. Chem. Phys., 10, 4725–4739, doi:10.5194/acp-10-4725-2010, 2010.

Zurita-Milla, R., van Gijsel, J. A. E., Hamm, N. A. S., Augustijn, P. W. M., and Vrieling, A.: Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps, IEEE T. Geosci. Remote, 51, 1914–1921, doi:10.1109/TGRS.2012.2223218, 2013.